Choosing a Search Engine

Federal Web Content Managers Workshop

April 27, 2005

David R. Baker

HHS Web Management Team

Excerpts from OMB Policy

- You must now ensure your agency’s principal public website and any major entry point include a search function. However, agencies may determine in limited circumstances (e.g., for small websites) site maps or subject indexes are more effective than a typical search function.

Excerpts from OMB Policy

- By December 31, 2005, this search function should, to the *extent practicable and necessary to achieve intended purposes*, permit searching of all files intended for public use on the website, display search results in order of relevancy to search criteria, and provide response times appropriately equivalent to industry best practices.

Excerpts from OMB Policy

- By December 31, 2005, agency public websites should to the extent practicable and necessary to achieve intended purposes, provide all *data in an open, industry standard format* permitting users to aggregate, disaggregate, or otherwise manipulate and analyze the data to meet their needs.

Excerpts from OMB Policy

- Agencies should note the Interagency Committee on Government Information has provided to OMB recommendations for organizing, categorizing, and searching for government information. By December 17, 2005, OMB will issue any necessary additional policies in this area.

HHS and Web Search

- HHS has a very diverse Web presence, with over 300 public Web sites containing several million pages.
- HHS organizations use many different search technologies on their sites.
- We needed a department-wide Web search, not an enterprise search.

HHS and Web Search

- HHS launched its new department-wide search on March 1st, after several months of testing.
- The search uses a Google Search Appliance.
- Five other components in HHS also have Google appliances.

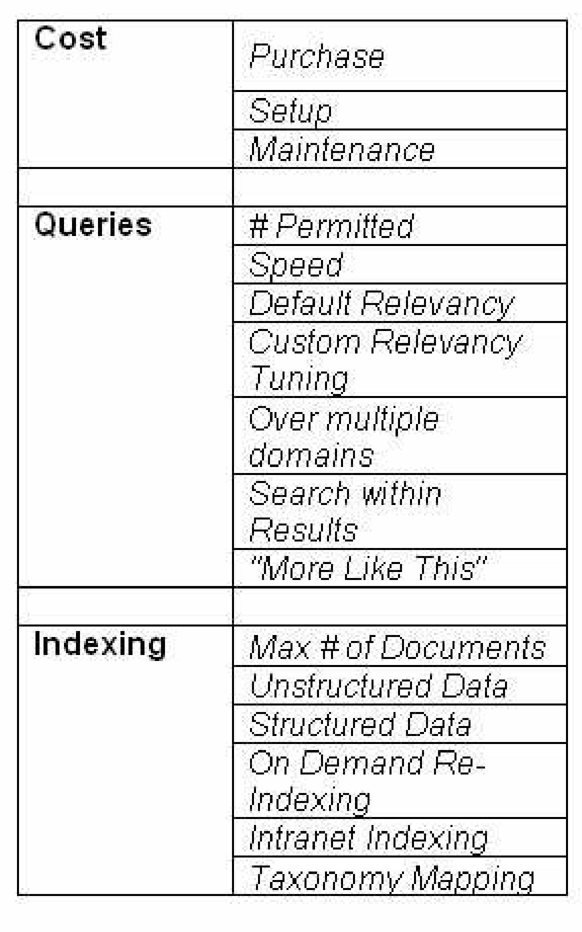
Selecting the Search Engine

- HHS began to review options for a new public Web search engine well before the OMB policy was issued.
- We formed a department-wide technical team to gather information and formulate draft requirements.
- Our portal implementation team then validated the requirements and handled the procurement.

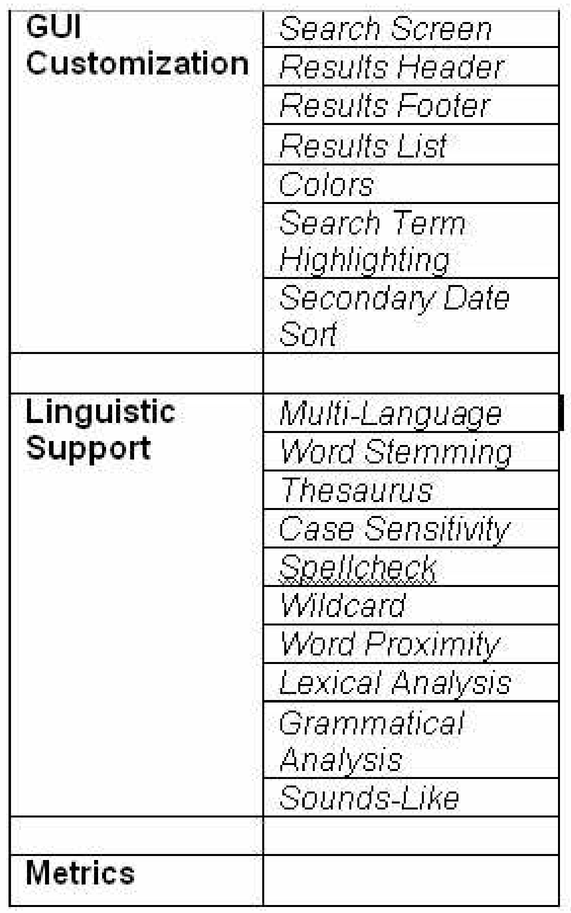
Selecting the Search Engine

- Both teams had HHS-wide representation.
- The teams also had practical experience with a broad range of search engines, including Google.
- That experience helped ground the requirements in the real world.

Considerations in Selection



Cost	*Purchase*
	Setup
	Maintenance
Queries	*# Permitted*
	Speed
	Default Relevancy
	Custom Relevancy Tuning
	Over multiple domains
	Search within Results
	"More Like This"
Indexing	*Max # of Documents*
	Unstructured Data
	Structured Data
	On Demand Re-Indexing
	Intranet Indexing
	Taxonomy Mapping



GUI Customization	*Search Screen*
	Results Header
	Results Footer
	Results List
	Colors
	Search Term Highlighting
	Secondary Date Sort
Linguistic Support	*Multi-Language*
	Word Stemming
	Thesaurus
	Case Sensitivity
	Spellcheck
	Wildcard
	Word Proximity
	Lexical Analysis
	Grammatical Analysis
	Sounds-Like
Metrics	

Considerations in Selection

- Our previous search was hosted, so new infrastructure and technical support requirements were important.
- Another key consideration was that our content was a mixed bag, with little metadata in place that could be leveraged to improve relevance of search results.
- The search engine and its relevance algorithm had to work in the real world.
- With government-wide search standards soon to be proposed, we didn’t want to be locked into a particular technology for a long period.

Search Strategy and Usage

- At this time, we are indexing only pages served as HTML pages.
- We exclude PDF, Microsoft Office, and other proprietary file formats to avoid confusing duplication in the search results.
- HHS posts enough documents in multiple formats that this was a concern.
- A high level of Section 508 compliance means almost all documents can be located through the search.

Search Strategy and Usage

- Our index includes about 725,000 HTML pages on 316 sites.
- This represents almost 3 times as many sites as we had in our old search.
- Users are performing about 200,000 searches per month, with the busiest day about 10,500 in April and the busiest hour just over 1,000.

Improvements in New Search

- Relevance of search results.
- Familiar user interface for the public due to Google’s status as one of the top three search engines.
- Timeliness—overnight turnaround for indexing new content.
- Spellchecking driven by our content.

Improvements in New Search

- Increased keyword control to match URLs to specific search terms and manage synonyms.
- Ability to *exclude* content from the index with greater specificity, reducing duplication.
- Can add keyword matches or remove URLs from search results in real time.

Improvements in New Search

- Ability to monitor search performance in real time.
- More timely reporting of search metrics.
- Ability to index content inside the firewall for our intranet.

Getting It Up and Running

- Turnkey solution. The Google appliance was an all-inclusive package of hardware and software, delivered and installed by a Google engineer.
- Instant availability. The appliance was up and running the same day.
- Reliability of a clustered solution.

Best Value

- Google’s search technology met our real-world requirements.
- Cost was predictable—a fixed price for hardware, software, and support for 2 years.
- The turnkey solution, including hardware replacement, minimized risk with regard to infrastructure and technical expertise.

Basic Search Engine Optimization

- Create usable, readable pages for people because search engine algorithms calculate relevance from a human perspective.
- Ensure navigation allows crawlers to reach all parts of your site.
- Include title, description, and keywords meta tags in the HTML header.
- Use keywords in meaningful headings.
- Use keywords at the beginning of the page in text.
- Use keywords in the URL or filename, and don’t change either unnecessarily.

Basic Search Engine Optimization

- Create alt tags for graphics containing keywords.
- Don’t try to spam a search engine by overuse or hidden use of keywords.
- Validate all HTML to ensure it can be ‘seen’ by the search engine.
- Use robots.txt and robots meta tags to keep search engines from indexing what they shouldn’t.
- Be sure your webserver supports the If-Modified-Since HTTP header.
- Avoid using frames.
- Avoid putting content and links within script code.

Contact

David R. Baker

david.baker@hhs.gov

202-260-1306